Introduction to Spark: Takeaways

Syntax

- Loading a data set into a resilient distributed data set (RDD):

```
raw_data = sc.textFile("daily_show.tsv")
```

- Printing out the first five elements of the RDD:

```
raw_data.take(5)
```

- Mapping a function to every element in the RDD:

```
daily_show = raw_data.map(lambda line: line.split('\t'))
```

- Merging the values for each similar key:

```
tally = daily_show.map(lambda x: (x[0], 1)).reduceByKey(lambda x,y: x+y)
```

- Retuning the count of an RDD:

```
tally.take(tally.count())
```

- Filtering an RDD:

```
rdd.filter(lambda x: x % 2 == 0)
```

Concepts

- MapReduce is a paradigm that efficiently distributes calculations over hundreds or thousands of computers to calculate the result in parallel.
- Hadoop is an open source project that is the primary processing tool kit for big data. There are pros and cons to Hadoop:
    - Hadoop made it possible to analyze large data sets; however, it had to rely on disk storage for computation rather than memory.
    - Hadoop wasn't a great solution for calculations that require multiple passes over the same data or require many intermediate steps.
    - Hadoop had suboptimal support for SQL and machine learning implementations.
- To improve the speeds of many data processing workloads, UC Berkeley AMP lab developed Spark.
- Spark's main core data structure is an RDD.
    - An RDD is a data set distributed across the RAM of a cluster of machines.
    - An RDD is essentially a collection of elements we can use to hold objects.
- PySpark is a Python API that allows us to interface with RDDs in Python.
- Spark's RDD implementation lets us evaluate code "lazily," which means we can postpone running a calculation until absolutely necessary.
- Calculations in Spark are a series of steps we can chain together and run in succession to form a pipeline. Pipelining is the key idea to understand when working with Spark.
- These are the main types of methods in Spark:
    - Transformations: `map()` , `reduceByKey()` .

    - Actions: `take()`, `reduce()`, `saveAsTextFile()`, `collect()`.
- RDD objects are immutable, which means that we can't change their values after we create them.

Resources

- MapReduce
- PySpark documentation